GLÀFF, un Gros Lexique À tout Faire du Français

Franck Sajous, Nabil Hathout, Basilio Calderone

GLÀFF, un Gros Lexique À tout Faire du Français

Franck Sajous, Nabil Hathout et Basilio Calderone
CLLE-ERSS – CNRS et Université de Toulouse 2 Le Mirail
{franck.sajous,nabil.hathout,basilio.calderone}@univ-tlse2.fr

RÉSUMÉ

Cet article présente GLÀFF, un lexique du français à large couverture extrait du Wiktionnaire, le dictionnaire collaboratif en ligne. GLÀFF contient pour chaque entrée une description morphosyntaxique et une transcription phonémique. Il se distingue des autres lexiques existants principalement par sa taille, sa licence libre et la possibilité de le faire évoluer de façon constante. Nous décrivons ici comment nous l'avons construit, puis caractérisé en le comparant à différentes ressources connues. Cette comparaison montre que sa taille et sa qualité font de GLÀFF un candidat sérieux comme nouvelle ressource standard pour le TAL, la linguistique et la psycholinguistique.

ABSTRACT

GLÀFF, a Large Versatile French Lexicon

This paper introduces GLÀFF, a large-scale versatile French lexicon extracted from Wiktionary, the collaborative online dictionary. GLÀFF contains, for each entry, a morphosyntactic description and a phonetic transcription. It distinguishes itself from the other available lexicons mainly by its size, its potential for constant updating and its copylefted license that makes it available for use, modification and redistribution. We explain how we have built GLÀFF and compare it to other known resources. We show that its size and quality are strong assets that could allow GLÀFF to become a reference lexicon for NLP, linguistics and psycholinguistics.

MOTS-CLÉS : Lexique morpho-phonologique, ressources lexicales libres, Wiktionnaire.

KEYWORDS: Morpho-phonological lexicon, free lexical resources, French Wiktionary.

1 Introduction

Pour être complet, il aurait fallu indiquer dans le nom de notre lexique qu'il est issu du Wiktionnaire, un très gros dictionnaire en ligne qui comporte plus de 2 millions d'entrées. À titre de comparaison, la nomenclature du *Trésor de la Langue Française* (TLF) contient environ 65 000 vedettes et 100 000 entrées (principales ou secondaires). À la taille du Wiktionnaire s'ajoute la grande variété de ses descriptions avec, outre les définitions, des informations relatives à la prononciation, aux formes fléchies, aux dérivés et aux membres de la famille morphologique, aux traductions, aux synonymes, antonymes, hyponymes, hyperonymes, etc. Ces informations paraissent répondre à une grande variété de besoins et être d'une qualité notable pour l'édition française. Notre projet est de permettre au TAL et plus généralement aux linguistes expérimentaux d'exploiter facilement cette ressource multi-usages, d'une richesse remarquable.

Une première exploitation du Wiktionnaire avait conduit à WiktionaryX [1] (Sajous *et al.*, 2010, 2011), un lexique structuré donnant accès, pour chaque entrée, à des informations de nature sémantique : définitions, synonymes, hyponymes, traductions dans d'autres langues, etc.
Avec GLÀFF [2], nous proposons une étape supplémentaire dont le but est de permettre l'exploitation des informations phonologiques et morphosyntaxiques. Outre sa taille et sa polyvalence, GLÀFF se caractérise par sa licence libre (Creative Commons By-SA). L'un des objectifs de ce travail est d'estimer la qualité de la ressource, sa couverture et son adéquation aux besoins des linguistes qui réalisent des expérimentations et/ou des modélisations, mais aussi des chercheurs et des développeurs de systèmes de TAL.

La suite de l'article s'organise comme suit : nous présentons dans la section 2 quelques-unes des ressources auxquelles GLÀFF peut être comparé. La section 3 présente le Wiktionnaire et différents travaux visant à construire à partir de ce dictionnaire des ressources pour le TAL. La construction de GLÀFF proprement dite fait l'objet de la section 4. Nous présentons en section 5 une série de comparaisons de GLÀFF avec des ressources de référence afin de quantifier les points forts et les apports de ce lexique. Plus précisément, nous nous intéressons à la couverture de GLÀFF en le comparant d'une part à quatre autres lexiques morphosyntaxiques disponibles du français, et en le projetant d'autre part sur différents corpus afin, notamment, de déterminer la proportion d'entrées attestées et non attestées. Nous comparons ensuite les descriptions phonémiques de GLÀFF avec celles de deux ressources qui en fournissent. Enfin, la section 6 conclut l'article et présente les étapes à venir dans le développement et l'exploitation de GLÀFF.

2 Ressources et travaux connexes

Des ressources lexicales pour le français commencent à être disponibles, même s'il reste beaucoup à faire pour nous rapprocher de la situation de l'anglais, tant en volume qu'en qualité. Le constat est similaire pour les outils généralistes comme les analyseurs morphosyntaxiques, syntaxiques, morphologiques et les outils de phonétisation, qui dépendent directement de ces ressources. Depuis la fin des années 1990, quelques ressources destinées au traitement automatique du français sont distribuées gratuitement : le lexique de l'ABU [3] date de 1999, la première version de Lefff (Clément *et al.*, 2004) de 2003 et Morphalou (Romary *et al.*, 2004) de 2004. Auparavant, seules des ressources payantes, principalement distribuées par ELRA, étaient disponibles.

La constitution de lexiques pour le TAL et pour l'étude outillée du français trouve son origine dans les travaux menés au LADL autour de Maurice Gross (Courtois, 1990; Silberztein, 1990). Les premiers étaient destinés à l'exploration de corpus et l'annotation lexicale. Les lexiques morphosyntaxiques du français fournissent tous un ensemble d'informations communes : la forme orthographique du mot, son lemme, la partie du discours et les propriétés morphosyntaxiques (traits flexionnels). Notons que si ces ressources ont été d'abord utilisées pour le TAL, elles le sont aussi pour la description linguistique, notamment en morphologie (Hathout *et al.*, 2009).

ABU, le plus ancien des lexiques morphosyntaxiques distribué librement sur le Web, comporte environ 60 000 lemmes et 300 000 formes. Les tailles de Lefff et de Morphalou sont plus importantes : respectivement 500 000 et 525 000 entrées. Notons que Lefff fournit également une description des cadres de sous-catégorisation des lexèmes.

1. Wiktionary XMLisé, disponible à l'adresse : `http://redac.univ-tlse2.fr/lexiques/wiktionaryx.html`
2. GLÀFF est disponible à l'adresse : `http://redac.univ-tlse2.fr/lexiques/glaff.html`
3. ABU : la Bibliothèque Universelle. `http://abu.cnam.fr`

À côté des ressources développées par des linguistes informaticiens (Lefff sert notamment à la mise au point d'analyseurs syntaxiques basés sur la théorie LFG) et des lexicographes (Morphalou est la version XML du lexique TLFnome, issu de la nomenclature du TLF), on trouve Lexique (New, 2006), qui s'inscrit dans la lignée de Brulex, une ressource développée à la fin des années 1980 (Content *et al.*, 1990). Comme Brulex, Lexique a été créé par et pour les psycholinguistes. Ces ressources se distinguent des lexiques morphosyntaxiques généralistes par la plus grande richesse des informations fournies : outre la morphosyntaxe, leur description lexicale comporte une transcription phonémique, une segmentation en syllabes, des informations sur les homophones, les homographes, les voisins phonologiques et orthographiques, la fréquence des formes dans des corpus écrits, etc. En contrepartie, l'absence d'un grand nombre des formes fléchies de Lexique, tout comme Brulex (seules les formes les plus usuelles y sont décrites) et les fréquences fournies, calculées à partir des graphies (qui ne tiennent donc pas compte des attributs morphosyntaxiques) constituent une limite à leur utilisation dans des outils de TAL. Lexique et Brulex sont actuellement les seules ressources gratuites qui fournissent des transcriptions phonémiques et un découpage syllabique. Il existe d'autres ressources plus complètes qui contiennent ces informations, créées dans des laboratoires de recherche publique... mais elles sont payantes[4]. L'une des plus anciennes et la plus connue est BDLex (Pérennou et de Calmès, 1987), dont la taille est similaire à celle de Lefff. Citons également ILPho (Boula De Mareuil *et al.*, 2000), plus récente, créée en complétant les entrées du lexique morphosyntaxique Multext (Ide et Véronis, 1994) par des transcriptions phonémiques. Dans tous ces lexiques, les transcriptions phonémiques sont codées au moyen de caractères ASCII, en SAMPA ou dans un format similaire.

3 Wiktionnaire

Wiktionary, le « *compagnon lexical de Wikipédia* », est un dictionnaire multilingue libre et accessible en ligne. Lancé en 2003, ce projet lexicographique fait état, 10 ans plus tard, de plus de deux millions d'articles pour son édition française, le *Wiktionnaire*. Si son remplissage a bénéficié de l'import d'articles du *Dictionnaire de l'Académie Française* et, dans une moindre mesure, du *Littré*, le Wiktionnaire connaît aujourd'hui une croissance constante grâce à l'édition manuelle des contributeurs. Chaque article peut contenir des informations étymologiques, définitions, exemples, relations sémantiques, traductions, transcriptions phonémiques, etc. Si l'on considère la couverture moindre des ressources lexicales existantes et les licences contraignantes sous lesquelles sont placées certaines d'entre elles, la variété des informations contenues dans le Wiktionnaire, la taille de sa nomenclature et sa mise à disposition sous licence libre en font un candidat extrêmement prometteur pour la construction d'un lexique électronique du français. Néanmoins, depuis l'émergence du *crowdsourcing*, se pose la question de la qualité des informations contenues dans les wikis. L'absence de comité éditorial et le fait que les modifications de tout contributeur, quelle que soit sa compétence, soient publiées immédiatement génèrent une certaine méfiance. À l'opposé, l'effet de mode lié à la naissance de nouveaux paradigmes peut générer un enthousiasme par trop optimiste (cf. la polémique portant sur la qualité de Wikipédia, opposant (Giles, 2005) à l'encyclopédie Britannica (Encyclopaedia Britannica, 2006)).

Si Wikipédia a fait l'objet d'analyses dans plusieurs disciplines et a servi en TAL de source de données pour constituer notamment des corpus et des listes d'entités nommées, ainsi que de base de calcul de similarité sémantique entre documents (Gabrilovich et Markovitch, 2007), Wiktionary n'a commencé à retenir l'attention des chercheurs, à notre connaissance, qu'en 2008. Il a d'abord

4. Outre le prix élevé de ces ressources, le fait de ne pouvoir en redistribuer des « œuvres dérivées » constitue une limite à la portée des travaux de recherche qui les utilisent, notamment en empêchant la reproductibilité des expériences.

été utilisé par (Zesch *et al.*, 2008), comme Wikipédia, comme point de départ pour effectuer des calculs de similarité sémantique. La qualité des ressources construites collaborativement « par les foules » et celles construites par les experts a été comparée par (Zesch et Gurevych, 2010), toujours à travers une tâche de mesure de similarité sémantique fondée sur Wikipédia et Wiktionary. Cette étude, plus modérée que celle de Giles, a montré que les ressources fondées sur « la sagesse des foules » ne sont pas meilleures que celles fondées sur « la sagesse des linguistes », mais sont sérieusement compétitives. Elles dépassent même les ressources construites par les experts dans certains cas, notamment en terme de couverture. Cependant, l'étude portait sur l'utilisation de données dérivées de Wiktionary et non sur son contenu primaire.

Le potentiel du Wiktionnaire en tant que lexique électronique n'a été étudié qu'à partir de 2009 par (Navarro *et al.*, 2009) pour le français et l'anglais.L'intégration de l'édition portugaise de Wiktionary dans l'ontologie Onto.PT (Gonçalo Oliveira et Gomes, 2010) est décrite dans (Anton Pérez *et al.*, 2011). Citons également Dbnary (Sérasset, 2012), une ressource et un extracteur *open source* visant à extraire de Wiktionary un réseau multilingue. L'auteur précise que ce travail ne vise pas l'exhaustivité mais la conception d'un modèle simple permettant de représenter autant de données qu'il est possible d'extraire correctement, laissant de côté certaines structures pour faciliter cette extraction. Le graphe extrait possède 260 467 entrées pour le français. Le laboratoire UKP distribue deux ressources issues de Wiktionary : OntoWiktionary (Meyer et Gurevych, 2012), une ontologie construite semi-automatiquement et UBY (Gurevych *et al.*, 2012), un alignement de 7 ressources lexicales incluant notamment WordNet, disponible pour l'allemand et l'anglais. Si la version allemande de Wiktionary semble être celle qui bénéficie de l'encodage le plus rigoureux et le plus systématique (par exemple, l'alignement avec d'autres ressources est permis par l'ancrage des relations sémantiques au niveau des sens, ce qui n'est pas le cas dans les éditions française et anglaise), la version française, moins aisément exploitable, se distingue par une plus grande nomenclature, ainsi que la présence quasi-systématique d'informations flexionnelles et phonémiques. Nous avons mis à disposition pour le français et l'anglais une version structurée au format XML de ce lexique. Nous présentons dans la section suivante l'extraction des informations phonémiques et morphosyntaxiques absentes de cette première version.

4 Construction

Pour chaque édition de langue, une mise à disposition régulière de l'ensemble des articles de Wiktionary est effectuée dans des fichiers appelés *XML dumps* [5]. Il ne faut pas interpréter la mention « XML » comme la structuration du contenu des articles par des balises qui délimiteraient les sections relatives aux catégories syntaxiques, relations sémantiques, traductions, etc. Les balises XML ne servent qu'à délimiter les articles et leur titre. Le reste du contenu est encodé dans un format appelé *wikicode*, inhérent au système de gestion de contenu *MediaWiki*. La syntaxe de ce format n'a jamais été définie formellement et, de plus, évolue dans le temps, avec coexistence de plusieurs conventions d'encodage pour un même type d'information. Il faut également mentionner que ni les conventions d'organisation des articles, ni leur encodage en *wikicode* n'est stable d'une édition de langue à l'autre. Nous montrons dans (Navarro *et al.*, 2009; Sajous *et al.*, 2010, 2011) comment ce format lâche rend ardue et constamment inachevée l'écriture d'un parseur pour extraire de manière automatique et exhaustive les informations de Wiktionary : entre la mise à disposition de deux *dumps*, le wikicode évolue sans que le changement ne soit nécessairement documenté et seule l'observation (semi-)manuelle du format d'encodage permet d'adapter le parseur en conséquence. Nous avons concentré notre effort dans

5. Voir : `http://dumps.wikimedia.org/`. Le dump utilisé pour ce travail est celui du 27/08/2012.

le travail présenté ici sur l'extraction des informations absentes de WiktionaryX : les informations flexionnelles et les transcriptions phonémiques.

La figure 1 montre un extrait de l'article *« affluent »*, tel qu'on peut le consulter dans le Wiktionnaire, deux de ses formes fléchies et le wikicode correspondant [6]. Le tableau qui recense les formes fléchies de l'adjectif, par exemple (en haut à droite de la figure 1a), n'est pas explicitement présent dans le wikicode, mais il est généré par le patron `{{fr-accord-cons|a.fly.ɑ̃|t}}`. Il existe ainsi des dizaines de patrons similaires dans le wikicode. L'extraction des formes fléchies et des prononciations correspondantes se fait soit par recensement et « émulation » de ces patrons (ici, génération des formes fléchies à partir d'un schéma spécifié), soit par l'analyse des articles des formes fléchies lorsqu'ils existent (cf. fig. 1c et 1d). Là encore, aucun formatage n'est systématique : le patron `{{f}}` (fig. 1c) indique que la forme est de genre féminin ; le nombre doit être extrait du texte de la ligne suivante *« Féminin singulier »*. Si la prononciation de *affluente* est donnée dans la *« ligne de forme »*, celle de *affluentes* est donnée dans une section *Prononciation* dédiée. Des erreurs induites par l'hétérogénéité du wikicode peuvent de ce fait s'ajouter aux erreurs contenues dans les articles du Wiktionnaire et ainsi impacter la ressource finale.

Notre parseur extrait du *dump* du Wiktionnaire les formes graphiques et leurs lemmes, convertit leurs catégories morphosyntaxiques au format GRACE (Rajman *et al.*, 1997) et extrait leurs transcriptions phonémiques, déjà en API. Notons qu'une même entrée peut avoir plusieurs transcriptions, comme *abricots*, dont on trouve une prononciation avec un « o » ouvert et une autre avec un « o » fermé : /a.bʁi.kɔ/ et /a.bʁi.ko/. Dans ce cas, toutes sont conservées.
Les informations flexionnelles présentes dans le Wiktionnaire sont parfois partielles. Il est en effet courant que seul le genre ou le nombre soit indiqué pour les noms et les adjectifs. De même, le temps ou le mode d'une forme verbale fléchie peut être omis. Nous appliquons des règles pour tenter de compléter ces informations : une forme nominale ou adjectivale ne portant pas de terminaison *-s* ou *-x*, par exemple, sera considérée comme étant au singulier ; le genre et le nombre d'un participe passé peuvent être inférés par sa terminaison ; une forme fléchie nominale ou adjectivale masculine, dont on a déjà rencontré le lemme masculin singulier, est plurielle ; etc. Les 9,5% d'entrées dont l'information flexionnelle reste partielle sont écartées de la ressource.
Dans cette première version de GLÀFF, dont un extrait est donné figure 2, ne sont inclus que les noms communs, verbes, adjectifs et adverbes (lemmes et formes fléchies). Les mots grammaticaux et locutions y seront intégrés dans les versions ultérieures. En complément du *dump* dont elles sont absentes, nous avons « aspiré » du site du Wiktionnaire, puis analysé, les tables de conjugaison de 18 076 verbes. Ces tables générées à partir d'un simple modèle (e.g. `{{fr-conj-1|march|pron=maʁ|pc=ʃ}}` pour le verbe *marcher* [7]) permettent d'obtenir les 48 flexions d'un verbe (nous n'intégrons pas les temps composés dans GLÀFF).

5 Caractérisation quantitative

La suite de l'article est consacrée à la caractérisation essentiellement quantitative de GLÀFF. Elle vise à apporter des éléments de réponse aux questions suivantes : que contient GLÀFF ? Quel est l'apport de GLÀFF relativement aux ressources similaires existantes ? GLÀFF est-il une ressource susceptible de remplacer les lexiques morphosyntaxiques et phonologiques courants ?
Cette caractérisation porte sur différents attributs : nombre de lemmes et de formes, couverture relativement à différents corpus et transcriptions phonémiques. Nous comparons GLÀFF à quatre

6. Un « guide pratique de parsing du wikicode » accompagnera prochainement la ressource GLÀFF.
7. Voir `http://fr.wiktionary.org/wiki/Annexe:Conjugaison_en_français/marcher`

affluent

Adjectif

affluent

1. *(Géographie)* Qui se jette dans un autre en parlant d'un cours d'eau.
2. *(Médecine)* Qui afflue, qui se porte en abondance vers quelque partie du corps.

	Singulier	Pluriel
Masculin	affluent /a.fly.ɑ̃/	affluents /a.fly.ɑ̃/
Féminin	affluente /a.fly.ɑ̃t/	affluentes /a.fly.ɑ̃t/

Nom commun

affluent /a.fly.ɑ̃/ *masculin*

1. *(Géographie)* Cours d'eau qui se jette dans un autre.

Singulier	Pluriel
affluent	affluents
/a.fly.ɑ̃/	

Forme de verbe

affluent /a.fly/

1. *Troisième personne du pluriel de l'indicatif présent de* affluer.
2. *Troisième personne du pluriel du subjonctif présent de* affluer.

Conjugaison du verbe *affluer*		
Indicatif	Présent	ils/elles affluent
Subjonctif	Présent	qu'ils/elles affluent

Prononciation

Adjectif et nom commun

- France : écouter « un affluent [ɛ̃.n‿a.fly.ɑ̃] »

(a) Mise en page de l'article « *affluent* »

```
{{-adj-|fr}}
{{fr-accord-cons|a.fly.ɑ̃|t}}
'''affluent'''
# {{géographie|fr}} Qui se [[jeter|jette]] [[dans]] un [[autre]] en [[parlant]] d'un [[cours]] d'eau.

{{-nom-|fr}}
{{fr-rég|a.fly.ɑ̃}}

{{-flex-verb-|fr}}
{{fr-verbe-flexion|affluer|ind.p.3p=oui|sub.p.3p=oui|}}
'''affluent''' {{pron|a.fly|fr}}
# ''Troisième personne du pluriel de l'indicatif présent de'' [[affluer]].
# ''Troisième personne du pluriel du subjonctif présent de'' [[affluer]].

{{-pron-}}
{| class="wikitable"
| Adjectif et nom commun
* {{pron-rég|France|ɛ̃ .n‿ a.fly.ɑ̃|titre=un affluent}}
|-
| Forme du verber affluer
* {{pron-rég|France (Île-de-France)|a.fly}}
* {{pron-rég|France (Île-de-France)|il.z‿ a.fly vɛʁ lɔ̃.tʁe dy ma.ga.zɛ̃|titre=ils
affluent vers l'entrée du magasin}} |}
```

(b) Wikicode de l'article « *affluent* »

affluente

Forme d'adjectif

affluente *féminin* /a.fly.ɑ̃t/

1. *Féminin singulier de* affluent.

```
{{-flex-adj-|fr}}
'''affluente''' {{f}} {{pron|a.fly.ɑ̃t|lang=fr}}
#''Féminin singulier de'' [[affluent#fr-adj|affluent]].
```

(c) Article « *affluente* » et wikicode correspondant

affluentes

Forme d'adjectif

affluentes

1. Féminin pluriel d'**affluent**.

Prononciation

- /a.fly.ɑ̃t/

```
{{-flex-adj-|fr}}
'''affluentes'''
# Féminin pluriel d''''[[affluent]]''''.

{{-pron-}}
* {{pron|a.fly.ɑ̃t}}
```

(d) Article « *affluentes* » et wikicode correspondant

FIGURE 1 – Article « *affluent* » et formes fléchies dans le Wiktionnaire.

```
affluent|Ncms|affluent|a.fly.ɑ̃
affluents|Afpmp|affluent|a.fly.ɑ̃
affluents|Ncmp|affluent|a.fly.ɑ̃
affluent|Vmip3p-|affluer|a.fly
affluent|Vmsp3p-|affluer|a.fly
afflueraient|Vmcp3p-|affluer|a.fly.ʁɛ
glênons|Vmmp1p-|glêner|gle.nɔ̃
glanure|Ncfs|glanure|gla.nyʁ
glanures|Ncfp|glanure|gla.nyʁ
glaoui|Ncms|glaoui|gla.wi
glapîmes|Vmis1p-|glapir|gla.pim
glapîtes|Vmis2p-|glapir|gla.pit
talentueuse|Afpfs|talentueux|ta.lɑ̃.tɥøz
talentueusement|Rgp|talentueusement|ta.lɑ̃.tɥøz.mɑ̃
talentueuses|Afpfp|talentueux|ta.lɑ̃.ty.øz
talent|Vmip3p-|taler|tal
talent|Vmsp3p-|taler|tal
taleraient|Vmcp3p-|taler|ta.lə.ʁɛ
```

FIGURE 2 – Extraits de GLÀFF

lexiques utilisés dans de nombreuses recherches : Lexique, BDLex, Morphalou et Lefff. Tous fournissent des descriptions morphosyntaxiques complètes pour leurs entrées, les deux premiers fournissant en plus des transcriptions phonémiques et une segmentation en syllabes.

Couverture. GLÀFF se distingue des lexiques actuellement utilisés en TAL et en psycholinguistique par sa taille exceptionnelle. La table 1 présente le nombre de lemmes et de formes fléchies, simples (séquence de lettres exclusivement) et non simples (*i.e.* comprenant espace, tiret et/ou chiffre). On peut y observer que GLÀFF contient 3 à 4 fois plus de lexèmes (2 fois plus pour les lemmes qui comportent une transcription phonémique) et 3 à 9 fois plus de formes (2 à 8 fois pour les formes transcrites). Cette taille est un atout important dans le cas d'une utilisation, par exemple, pour des recherches en morphologie flexionnelle ou dérivationnelle. Elle est également intéressante pour le développement d'outils de TAL comme des étiqueteurs morphosyntaxiques ou des analyseurs syntaxiques. On observe également que GLÀFF comporte un nombre élevé de formes composées. Ces dernières servent essentiellement à la segmentation des textes en « tokens » dont la qualité impacte l'ensemble des annotations catégorielles et syntaxiques ultérieures.

		Formes fléchies catégorisées			**Lemmes catégorisés**		
		Simples	Non simples	Total	Simples	Non simples	Total
Lexique		147 912	4 696	152 608	46 649	3 770	50 419
BDLex		431 992	4 360	436 352	47 314	1 792	49 106
Lefff		466 668	3 829	470 497	54 214	2 303	56 517
Morphalou		524 179	49	524 228	65 170	7	65 177
GLÀFF	Avec transcription	1 258 217	11 209	1 269 426	105 646	6 091	111 737
	Sans transcription	143 361	13 061	156 422	66 970	7 375	74 345
	Total	1 401 578	24 270	1 425 848	172 616	13 466	186 082

TABLE 1 – Taille des lexiques (restreints aux catégories : nom commun, verbe, adjectif, adverbe).

Les comparaisons ci-après concernent uniquement les catégories majeures nom commun, verbe, adjectif et adverbe. Elles ont été réalisées sur les formes graphiques ou lemmes « simples » afin de nous affranchir des différents choix de graphie des unités polylexicales dans les lexiques et de segmentation des corpus. Nous étudions tout d'abord l'intersection de GLÀFF avec les autres lexiques. La table 2 présente pour chacun des cinq lexiques testés la proportion d'entrées (*i.e.* de triplets *<forme ; lemme ; description morphosyntaxique>*) que l'on retrouve à l'identique dans les autres. On observe que la taille des intersections est directement liée à celle des lexiques : plus un lexique est gros, plus son intersection avec les autres l'est. On observe ensuite une répartition des cinq lexiques en trois groupes : Lexique a une couverture moindre, avec 9% de GLÀFF et 22 à 26% des lexiques BDLex, Lefff et Morphalou. Ces trois derniers couvrent 76% à 80% de Lexique et 30% de GLÀFF en moyenne, tout en ayant un couverture commune de 70% à 86%. GLÀFF est nettement au-dessus avec un couverture de 85% à 93%. Sa couverture est supérieure de 5% à 13% à celle des autres lexiques. De plus, le fait qu'il n'inclue que partiellement les autres lexiques est normal au vu des intersections de ces derniers.

	Lexique	**BDLex**	**Lefff**	**Morphalou**	**GLÀFF**
Lexique		26,03	25,20	22,46	8,95
BDLex	76.02		79,87	70,40	28,75
Lefff	79,50	86,28		72,32	30,04
Morphalou	79,58	85,43	81,24		32,03
GLÀFF	**84,83**	**93,26**	**90,23**	**85,66**	

TABLE 2 – Couverture inter-lexiques (en % de formes fléchies catégorisées).

GLÀFF a donc une taille nettement supérieure à celle des autres lexiques, ce qui constitue un atout potentiel. Afin de s'assurer que cet avantage est effectif (*i.e.* que le plus grand nombre de lexèmes et de formes peut réellement s'avérer utile), nous avons comparé les cinq lexiques au vocabulaire de quatre corpus de nature différente (genre, taille, époque, etc.). Le premier, composé de 515 romans du XX^e siècle issus de la base Frantext [8], contient 30 millions de mots. LM10, corpus journalistique qui rassemble les archives de 1991 à 2000 du quotidien *Le Monde*, contient 200 millions de mots. Le troisième corpus composé des 664 982 articles de la Wikipédia française [9], contient 260 millions de mots. Enfin, FrWaC (Baroni *et al.*, 2009) est un corpus de pages Web en français contenant 1,6 milliard de mots. Ces quatre corpus ont été étiquetés par la version standard de TreeTagger [10], qui nous sert ici à segmenter les corpus et filtrer leur vocabulaire sur la base des catégories syntaxiques (qui sont ensuite ignorées). Les mots inconnus de TreeTagger (dont la catégorie est pertinente) sont conservés. La table 3 présente la couverture des cinq lexiques par rapport à ces quatre corpus, en distinguant au sein de leur vocabulaire les formes de fréquence supérieure ou égale à 1 (*i.e.* tout le vocabulaire), 2, 5, 10, 100 et 1000.

Seuil : fréquence ≥		**1**	**2**	**5**	**10**	**100**	**1000**
Frantext	Nb formes	145 437	95 189	61 813	43 919	10 767	1 376
	Lexique	66,76	84,35	94,00	**96, 91**	**99, 15**	**99, 27**
	BDLex	70,86	84,69	92,47	95,74	99,12	99,20
	Lefff	71,89	85,63	93,21	96,21	99,08	98,90
	Morphalou	73,93	86,66	93,29	96,00	98,48	97,09
	GLÀFF	**76, 92**	**88, 57**	**94, 54**	96,72	98,77	98,76
LM10	Nb formes	300 606	172 036	106 470	77 936	29 388	7 838
	Lexique	29,59	47,28	65,23	76,31	93,81	98,58
	BDLex	37,77	55,79	71,76	80,93	95,53	98,69
	Lefff	39,64	58,22	74,33	83,20	95,99	**98, 90**
	Morphalou	39,06	56,82	71,92	80,32	93,27	97,48
	GLÀFF	**45, 24**	**63, 83**	**78, 63**	**86, 23**	**96, 46**	98,68
Wikipédia	Nb formes	953 920	435 031	216 210	136 531	35 621	7 956
	Lexique	9,13	18,27	31,52	43,03	78,58	95,72
	BDLex	12,29	22,89	36,80	48,04	79,39	95,33
	Lefff	12,88	23,94	38,26	49,65	80,57	95,71
	Morphalou	13,05	23,96	37,87	48,87	78,74	94,16
	GLÀFF	**16, 42**	**29, 00**	**44, 13**	**55, 45**	**83, 21**	**96, 10**
FrWaC	Nb formes	1 624 620	846 019	410 382	255 718	74 745	22 100
	Lexique	5,83	10,85	20,84	30,81	66,00	89,47
	BDLex	9,36	15,85	27,28	37,48	69,61	90,03
	Lefff	9,85	16,67	28,57	39,16	71,61	91,16
	Morphalou	10,09	16,89	28,53	38,68	69,36	88,51
	GLÀFF	**13, 13**	**21, 13**	**34, 29**	**45, 35**	**76, 39**	**92, 76**

TABLE 3 – Couverture lexiques/corpus (en % de formes fléchies non catégorisées).

Le classement des corpus par couverture décroissante est le même pour les cinq lexiques. Bien que la taille des corpus influe sur cet ordre (plus un corpus est étendu, plus le nombre potentiel de formes différentes est grand), leur nature est également déterminante : FrWaC, par exemple, est une collection de pages web et (donc) contient nombre de formes « bruitées » (mots étrangers, espaces manquants ou excédentaires, orthographe aléatoire, absence de diacritiques, etc.). On retrouve la répartition des lexiques en trois groupes : BDLex, Lefff et Morphalou présentent une couverture assez proche. Hormis pour Frantext, Lexique affiche une couverture moindre jusqu'au

8. `http://www.frantext.fr/`
9. Version du 18 juin 2008 disponible à l'adresse : `http://redac.univ-tlse2.fr/corpus/wikipedia.html`
10. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

seuil 100, où il rejoint Morphalou. GLÀFF a une couverture supérieure pour les trois plus gros corpus, sauf pour LM10 au seuil 1000 où il est dépassé par Lefff de 0,2%. La meilleure couverture de Lexique pour Frantext, alors qu'elle est inférieure de 10 à 15% à celle de GLÀFF pour les trois autre corpus, s'explique probablement par le fait que son vocabulaire a été constitué à partir d'œuvres de cette même base. Pour les autres corpus et jusqu'au seuil 100, la taille de GLÀFF lui permet d'avoir une couverture du vocabulaire bien supérieure à celle des autres lexiques (au seuil 1, de 14% à 53% de plus pour LM10 et de 30% à 125% pour FrWaC ; au seuil 10, de 4% à 16% pour LM10 et de 15% à 47% pour FrWaC). Des outils de TAL qui intégreraient GLÀFF devraient donc améliorer leurs performances dans le traitement de ces corpus.

La figure 3 compare la couverture des cinq lexiques sous un autre éclairage : elle représente pour chaque lexique le nombre de formes dont la fréquence en corpus appartient à un intervalle donné. On y voit clairement que les différences sont plus marquées pour le corpus FrWaC qu'elle ne le sont pour Frantext, probablement du fait des différences liées à la nature des corpus, comme expliqué plus haut. La répartition des lexiques en trois groupes apparaît clairement dans le diagramme de droite (FrWaC). On voit également sur ce dernier que même pour les mots très fréquents et donc très bien attestés, qui ont par exemple une fréquence comprise entre 101 et 1000, la couverture de GLÀFF reste meilleure. La table 3 et la figure 3 montrent que la supériorité de GLÀFF est plus marquée lorsque l'on travaille sur des corpus hétérogènes et pour des mots de faibles et moyennes fréquences.

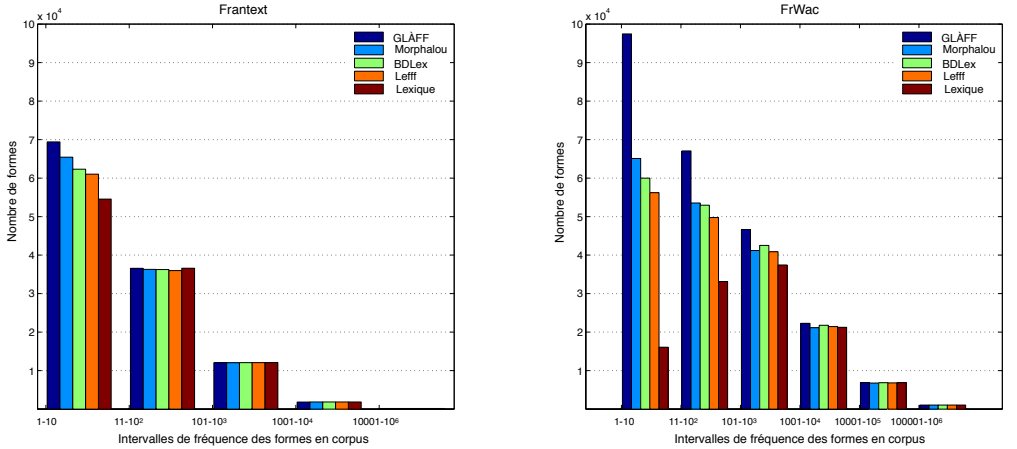

FIGURE 3 – Répartition des formes des lexiques relativement à leur fréquence en corpus.

Pour conclure notre caractérisation de la couverture de GLÀFF, nous nous sommes intéressés à la partie du vocabulaire spécifique à ce dernier, *i.e.* aux formes appartenant à GLÀFF et absentes des quatre autres lexiques. Ce sous-ensemble de 665 290 formes représente 47% de la ressource. Nous avons également considéré le vocabulaire spécifique de chaque autre lexique. La table 4 montre pour chaque sous-vocabulaire le nombre de formes attestées en corpus. Conformément à l'intuition, le nombre de formes attestées est d'autant plus grand que les corpus sont gros. La taille des corpus n'explique cependant pas tout : si une large part du vocabulaire spécifique à GLÀFF n'est attestée dans aucun corpus (il s'agit majoritairement de verbes pour lesquelles toutes les flexions possibles sont générées), sa taille permet une meilleure couverture de corpus hétérogènes tel que FrWaC incluant un français potentiellement moins normé et plus récent.

Même pour un corpus journalistique dont l'année la plus récente est 2000, la jeunesse, mais également la mise à jour constante du Wiktionnaire permettent à GLÀFF de couvrir des mots tout à fait usuels comme : *transversalité*, *attractivité*, *brevetabilité*, *diabolisation*, *employabilité*, *anticorruption*, *homophobie*, *institutionnellement*, *hébergeur*, *fatwa*, *indétrônable*, etc., toujours absents des autres lexiques après 13 années.

	Taille du vocabulaire spécifique	Nombre de formes attestées			
		Frantext	LM10	Wikipédia	FrWaC
Lexique	1 509	866	863	1 073	1 320
BDLex	3 981	86	521	1 004	1 496
Lefff	11 050	232	1 479	2 214	3 288
Morphalou	26 881	1 171	1 912	3 995	6 425
GLÀFF	665 290	2 811	13 525	29 230	47 549

TABLE 4 – Attestation en corpus du vocabulaire spécifique de chaque lexique

Transcriptions phonémiques. GLÀFF contient, pour 90% de ses entrées, une transcription phonémique. Ces transcriptions contiennent parfois (8% des cas) plusieurs variantes. Afin d'évaluer leur qualité, nous les avons comparées à celles de BDLex et Lexique, que nous avons converties en API. Nous avons comparé d'une part les transcriptions sans tenir compte de la syllabation, puis nous avons comparé la syllabation pour les transcriptions dont les suites de phonèmes sont strictement identiques. Nous avons relevé, pour les transcriptions qui ne diffèrent que par un phonème, les oppositions en cause. La table 5 montre pour chaque couple de lexiques les 10 oppositions les plus fréquentes et la table 6 donne des exemples illustrant ces oppositions. Cette table est complétée, dans la dernière colonne, par les transcriptions du *Dictionnaire de la Prononciation Française dans son Usage Réel* (Martinet et Walter, 1973), noté DPF ci-après. Les auteurs de ce dictionnaire papier élaboré entre 1968 et 1973, partant du principe que « *l'unité de la prononciation française est une vue de l'esprit et ne correspond à rien de réel* » ont mené, avec leurs collaborateurs, un travail de recensement auprès de 17 informateurs pour collecter les différentes variantes de prononciation d'un même mot (20% des prononciations divergent).
Les différences de transcription entre GLÀFF et chacun des deux autres lexiques sont comparables aux différences que l'on trouve entre BDlex et Lexique. Elles sont principalement dues à l'opposition entre les voyelles moyennes, comme les antérieures : [e] (mi-fermée) vs. [ɛ] (mi-ouverte), et les postérieures : [o] (mi-fermée) vs. [ɔ] (mi-ouverte). Entre BDLex et Lexique, elles sont responsables de 91% des divergences. Ces différences étaient attendues : l'opposition entre voyelles mi-fermeés et mi-ouvertes en français est soumise à des restrictions distributionnelles définies par la *loi de position* selon laquelle les voyelles mi-ouvertes apparaissent de préférence en syllabe fermée, alors que les voyelles mi-fermées apparaissent de préférence en syllabe ouverte. Bien qu'une investigation détaillée sur les structures syllabiques reste à faire, la variabilité d'application de cette loi n'est pas uniforme en France : elle est plus systématique pour le Midi, moins pour le Nord (Detey *et al.*, 2010). Les autres oppositions relevées dans la table 5, comme l'opposition [s]/[z], venant principalement du suffixe *-isme*, sont décrites dans le DPF. Le codage problématique du schwa y est également longuement commenté. La table 7 montre la proportion de transcriptions strictement identiques (hors syllabation), et « comparables » après annulation des différences entre voyelles moyennes. Notre définition de *comparable* est arbitraire mais montre que la majorité des différences (97 à 98%) sont dues à ces seules oppositions et ne viennent pas de codages aberrants. GLÀFF et Lexique proposent des prononciations strictement identiques pour 79,5% des entrées. Cet accord strict est de 61,7% entre GLÀFF est BDLex. On peut donc estimer que les transcriptions phonémiques de GLÀFF sont de bonne qualité (l'accord entre

BDLex et Lexique est de 58,3%). Notons également que les emprunts sont souvent générateurs de divergences (e.g. *shaker* : /ʃɛi.kəʁ/, /ʃɛj.kœʁ/, /ʃɛ.kœʁ/ ; *chili* : /ʃi.li/, /tʃi.li/ ; *ginseng* : /ʒin.sɑ̃g/, /ʒin.sɑ̃ŋ/, /ʒin.sɛŋ/). Par ailleurs, ni Lexique ni BDLex ne saurait constituer un étalon absolu. Si l'opposition [o]/[ɔ] peut s'expliquer, certaines entrées transcrites avec un [o] (o fermé) dans BDLex sont surprenantes : /po,m/ pour *pomme*, /poʁt/ pour *porte*, /oʁ/ pour *or* et *hors*, etc. Concernant Lexique, que penser de *châtié*, transcrit /ʃa.sje/, ou de *cambriolé/cambriolés* transcrits respectivement /kɑ̃.bʁi.jo.le/ et /kɑ̃.bʁi.o.le/ ? On s'étonne également de lire dans sa documentation [11] que le caractère 9 code le « e-ouvert [comme dans] *œuf, peur* » et de trouver dans le lexique *peur* transcrit /p2R/, 2 étant selon la documentation le code pour le « e-fermé [comme dans] *deux* ». Une autre curiosité concerne le « schwa non élidable [comme dans] *parvenu* », codée selon la documentation par le symbole 3. Or ce symbole est totalement absent de Lexique. *Parvenu*, utilisé comme exemple, est transcrit /paRv˚ny/, où ˚ code le schwa élidable.

Op.	Phonèmes	%	% cumulé
r	ɛ/e	48,18	48,18
r	ɔ/o	32,17	80,36
r	o/ɔ	11,02	91,37
r	y/ɥ	1,83	93,21
r	ə/ø	1,44	94,64
r	ə/œ	1,39	96,03
r	u/w	0,84	96,87
r	b/p	0,73	97,61
r	s/z	0,51	98,12
d	j	0,25	98,37

(a) BDLex/Lexique

Op.	Phonèmes	%	% cumulé
r	ɔ/o	60,03	60,03
i	ə	14,18	74,21
r	e/ɛ	6,90	81,11
r	ɛ/e	4,98	86,09
r	ɑ/a	4,92	91,01
r	s/z	1,25	92,26
r	ə/ø	0,91	93,17
r	œ/ø	0,47	93,64
i	i	0,42	94,06
r	o/ɔ	0,38	94,44

(b) GLÀFF/Lexique

Op.	Phonèmes	%	% cumulé
r	e/ɛ	66,46	66,46
r	ɔ/o	10,58	77,05
i	ə	5,90	82,96
r	o/ɔ	4,36	87,32
r	ɑ/a	3,84	91,17
r	ɥ/y	1,61	92,78
r	œ/ə	1,09	93,88
r	ø/ə	0,86	94,74
i	i	0,84	95,58
r	w/u	0,79	96,38

(c) GLÀFF/BDLex

TABLE 5 – Les 10 différences de transcription les plus fréquentes.
Opérations (Op.) : r = substitution ; i = insertion ; d = suppression.

		Transcriptions			
Opération	**Forme**	**BDLex**	**Lexique**	**GLÀFF**	**DPF**
r : ɛ/e	été	/ɛ.te/	/e.te/	/e.te/	/ete/
r : s/z	stalinisme	/sta.li.nis,m/	/sta.li.nizm/	/sta.li.nism/	/stalinism/, /stalinizm/
r : b/p	obturer	/ɔb.ty.ʁe/	/ɔp.ty.ʁe/	/ɔp.ty.ʁe/	/ɔptyre/, /ɔbtyre/
r : o/ɔ	pomme	/po,m/	/pɔm/	/pɔm/	/pɔm/
r : ə/ø/œ	heureux	/ə.ʁø/	/ø.ʁø/	/œ.ʁø/	/ørø, œrø/
r : y/ɥ	gradué	/gʁa.dy.e/	/gʁa.dɥe/	/gʁa.dɥe/	/gradɥe/, /grɑdɥe/, /gradye/
r : u/w	jouer inouï	/ʒu.e/ /i.nu.i/	/ʒwe/ /i.nwi/	/ʒwe/ /i.nwi/	/ʒwe/, /ʒue/ /inwi/, /inui/
r : a/ɑ	pâte	/pa,t/	/pat/	/pɑt/	/pat/ , /pɑt/
i,d : i,j	riiez	/ʁi.i.je/	/ʁi.je/	/ʁij.je/	-
i,d : ə	contenu	/kɔ̃,tə.ny/	/kɔ̃.tə.ny/	/kɔ̃t.ny/	/kɔ̃t(ə)ny/

TABLE 6 – Exemples de différence de transcription entre lexiques.

La comparaison de la syllabation opérée sur les transcriptions identiques (cf. table 7) montre que les trois lexiques sont très proches (98%). Notons à ce propos que si la construction « collaborative par les foules » du Wiktionnaire peut dans certains cas, admettons-le, être source d'amateurisme, elle peut également être intéressante car elle reflète une perception non canonique de la langue, selon un point de vue qui est *de facto* celui du locuteur (en l'occurrence, le contributeur) et non

11. http://www.lexique.org/outils/Manuel_Lexique.htm

de jure celui du linguiste. À titre d'exemple, on peut citer le cas de la syllabation du groupe consonantique /s/ + C en position interne de mot. Dans GLÀFF ce groupe apparaît alternativement comme hétérosyllabique, *i.e.* le /s/ et la consonne qui suit appartiennent à deux syllabes différentes (c'est la version canonique en français) comme dans *ministère* /mi.nis.tɛʁ/ et comme tautosyllabique (les deux phonèmes appartiennent à la même syllabe) comme dans *monistique* /mɔ.ni.stik/. Cette alternance, avec d'autres phénomènes non stables dans le Wiktionnaire, peuvent être perçus comme les signaux du comportement parfois non déterministe de la langue, et, partant, comme des objets potentiels d'investigation linguistique et psycholinguistique.

			Transcriptions phonémiques		Syllabation
Lexiques		Intersection	Identiques	Comparables	Identiques
BDLex	Lexique	112 439	58,31	96,88	98,92
GLÀFF	Lexique	123 630	79,50	97,81	98,48
GLÀFF	BDLex	396 114	61,72	96,88	98,30

TABLE 7 – Accord inter-lexiques : transcriptions phonémiques et syllabation
(Transcriptions comparables : non prise en compte des oppositions [o]/[ɔ], [e]/[ɛ] et [œ]/[ə]/[ø].
Syllabation : comparaison sur les transcriptions phonémiques identiques)

6 Conclusion et perspectives

Nous avons présenté dans cet article la première version d'un nouveau lexique, GLÀFF, construit de façon automatique à partir du Wiktionnaire. Ce lexique fournit des descriptions morphosyntaxiques détaillées pour 1,4 millions d'entrées et des transcriptions phonémiques pour 1,3 millions d'entre elles. Nous avons apporté dans cet article un certain nombre d'éléments qui indiquent que GLÀFF est un lexique de bonne qualité comparé aux ressources existantes comme Lexique, BDLex, Lefff ou Morphalou. Le fait qu'il dispose d'une taille 3 à 9 fois supérieure à celle des autres lexiques ne s'accompagne pas d'une dégradation des descriptions morphosyntaxiques et des transcriptions phonémiques. GLÀFF devrait s'avérer utile tant pour des recherches en TAL, en psycholinguistique, que pour la description linguistique.

La création de GLÀFF, motivée notamment par les besoins des recherches que nous menons sur l'organisation morphologique du lexique (Hathout, 2011) et sur la modélisation de la phonotactique et de son acquisition (Calderone et Celata, 2012), se poursuivra par un travail sur la découverte automatique des espaces thématiques utilisés pour la flexion (Boyé, 2011). Les autres perspectives sont nombreuses. À très court terme, nous enrichirons GLÀFF des catégories syntaxiques initialement écartées (mots grammaticaux et locutions). Puis nous intégrerons dans une même ressource les informations contenues dans GLÀFF et WiktionaryX. Cette ressource unifiée pourra dans un second temps recevoir des bases de données de descriptions lexicales provenant par exemple du *Dictionnaire des Mots Construits* de Michel Roché [12].

Nous prévoyons également la création d'une version révisée de GLÀFF qui constituera un sous-lexique totalement fiable. Plusieurs stratégies sont envisagées. La première sera de détecter automatiquement les entrées susceptibles de comporter des erreurs dans leur description morphosyntaxique ou leur transcription phonémique et de les éliminer. La seconde sera une révision semi-automatique dans laquelle nous proposerons à des opérateurs humains des corrections possibles qu'ils devront valider. Nous pourrons enfin augmenter ces sous-lexiques par une collection étendue d'informations sur la fréquence des formes et des lexèmes dans différents corpus

12. http://w3.erss.univ-tlse2.fr/textes/pagespersos/mroche/

de référence (Frantext, FrWaC, Wikipédia, LM10, etc.), le nombre de caractères, de phonèmes, de syllabes, la taille de la famille dérivationnelle, le voisinage graphémique, phonologique, etc. Cette version devrait répondre aux besoins des psycholinguistes, et être également utile pour la description linguistique, notamment en morphologie, les études quantitatives en linguistique et la modélisation du lexique et de son acquisition.

Références

ANTON PÉREZ, L., GONÇALO OLIVEIRA, H. et GOMES, P. (2011). Extracting Lexical-Semantic Knowledge from the Portuguese Wiktionary. *In Proceedings of the 15th Portuguese Conference on Artificial Intelligence*, EPIA 2011, pages 703–717. APPIA.

BARONI, M., BERNARDINI, S., FERRARESI, A. et ZANCHETTA, E. (2009). The WaCky wide web : a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

BOULA DE MAREUIL, P., YVON, F., D'ALESSANDRO, C., AUBERGÉ, V., VAISSIÈRE, J. et AMELOT, A. (2000). A French Phonetic Lexicon with variants for Speech and Language Processing. *In Proc. of the 2nd Intl Conference on Language Resources and Evaluation (LREC)*, pages 273–276.

BOYÉ, G. (2011). Régularité et classes flexionnelles dans la conjugaison du français. *In* (Roché *et al.*, 2011), pages 41–68.

CALDERONE, B. et CELATA, C. (2012). PHACTS about activation-based word similarity effects. *In Proceedings of the EACL 2012 Workshop on Computational Models of Language Acquisition and Loss*, pages 33–37, Avignon. ACL.

CLÉMENT, L., LANG, B. et SAGOT, B. (2004). Morphology based automatic acquisition of large-coverage lexica. *In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1841–1844, Lisboa, Portugal.

CONTENT, A., MOUSTY, P. et RADEAU, M. (1990). BRULEX : Une base de données lexicales informatisée pour le français écrit et parlé. *L'Année Psychologique*, 90:551–566.

COURTOIS, B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue française*, 87(1):11–22.

DETEY, S., DURAND, J., LAKS, B. et LYCHE, C. (2010). *Les variétés du français parlé dans l'espace francophone*. L'essentiel francais. Ophrys.

ENCYCLOPAEDIA BRITANNICA (2006). Fatally Flawed : Refuting the Recent Study on Encyclopedic Accuracy by the Journal Nature.

GABRILOVICH, E. et MARKOVITCH, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611.

GILES, J. (2005). Internet Encyclopaedias go Head to Head. *Nature*, 438:900–901.

GONÇALO OLIVEIRA, H. et GOMES, P. (2010). Onto.PT : Automatic Construction of a Lexical Ontology for Portuguese. *In Proceedings of 5th European Starting AI Researcher Symposium*, pages 199–211. IOS Press.

GUREVYCH, I., ECKLE-KOHLER, J., HARTMANN, S., MATUSCHEK, M., MEYER, C. M. et WIRTH, C. (2012). UBY - A Large-Scale Unified Lexical-Semantic Resource Based on LMF. *In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 580–590.

HATHOUT, N. (2011). Une approche topologique de la construction des mots : propositions théoriques et application à la préfixation en *anti-*. *In* (Roché *et al.*, 2011), pages 251–318.

HATHOUT, N., NAMER, F., PLÉNAT, M. et TANGUY, L. (2009). La collecte et l'utilisation des données en morphologie. *In* FRADIN, B., KERLEROUX, F. et PLÉNAT, M., éditeurs : *Aperçus de morphologie du français*, pages 267–287. Presses universitaires de Vincennes, Saint-Denis.

IDE, N. et VÉRONIS, J. (1994). MULTEXT : Multilingual text tools and corpora. *In Proceedings of the 15th conference on Computational linguistics (COLING94)*, pages 588–592, Kyoto, Japan.

MARTINET, A. et WALTER, H. (1973). *Dictionnaire de la Prononciation Française dans son Usage Réel*. France Expansion.

MEYER, C. M. et GUREVYCH, I. (2012). OntoWiktionary – Constructing an Ontology from the Collaborative Online Dictionary Wiktionary. *In* PAZIENZA, M. T. et STELLATO, A., éditeurs : *Semi-Automatic Ontology Development : Processes and Resources*, chapitre 6, pages 131–161. IGI Global, Hershey, PA, USA.

NAVARRO, E., SAJOUS, F., GAUME, B., PRÉVOT, L., HSIEH, S., KUO, I., MAGISTRY, P. et HUANG, C.-R. (2009). Wiktionary and NLP : Improving synonymy networks. *In Proceedings of the 2009 ACL-IJCNLP Workshop on The People's Web Meets NLP : Collaboratively Constructed Semantic Resources*, pages 19–27, Suntec, Singapore. Association for Computational Linguistics.

NEW, B. (2006). Lexique 3 : Une nouvelle base de données lexicales. *In Verbum ex machina. Actes de la 13e conférence sur le Traitement automatique des langues naturelles*, Louvain-la-Neuve.

PÉRENNOU, G. et de CALMÈS, M. (1987). BDLEX lexical data and knowledge base of spoken and written French. *In Proceedings of the European Conference on Speech Technology, ECST 1987*, pages 1393–1396, Edinburgh, Scotland, UK.

RAJMAN, M., LECOMTE, J. et PAROUBEK, P. (1997). Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique. Rapport technique, EPFL & INaLF. GRACE GTR-3-2.1.

ROCHÉ, M., BOYÉ, G., HATHOUT, N., LIGNON, S. et PLÉNAT, M. (2011). *Des unités morphologiques au lexique*. Hermès Science-Lavoisier, Paris.

ROMARY, L., SALMON-ALT, S. et FRANCOPOULO, G. (2004). Standards going concrete : from LMF to Morphalou. *In* ZOCK, M. et SAINT-DIZIER, P., éditeurs : *COLING 2004 Enhancing and using electronic dictionaries*, pages 22–28, Geneva. COLING.

SAJOUS, F., NAVARRO, E. et GAUME, B. (2011). Enrichissement de lexiques sémantiques approvisionnés par les foules : le système WISIGOTH appliqué à Wiktionary. *TAL*, 52(1):11–35.

SAJOUS, F., NAVARRO, E., GAUME, B., PRÉVOT, L. et CHUDY, Y. (2010). Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources : Piggybacking onto Wiktionary. *In* LOFTSSON, H., RÖGNVALDSSON, E. et HELGADÓTTIR, S., éditeurs : *Advances in Natural Language Processing*, volume 6233 de *LNCS*, pages 332–344. Springer Berlin / Heidelberg.

SILBERZTEIN, M. (1990). Le dictionnaire électronique des mots composés. *Langue française*, 87(1):71–83.

SÉRASSET, G. (2012). Dbnary : Wiktionary as a LMF based Multilingual RDF network. *In Proc. of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul.

ZESCH, T. et GUREVYCH, I. (2010). Wisdom of Crowds versus Wisdom of Linguists - Measuring the Semantic Relatedness of Words. *Journal of Natural Language Engineering.*, 16(01):25–59.

ZESCH, T., MÜLLER, C. et GUREVYCH, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. *In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.